

Listener Acoustic Personalization Challenge - LAP24: Head-Related Transfer Function Upsampling

AIDAN O. T. HOGG ^{1,2}, ROBERTO BARUMERLI ³, RAPOLAS DAUGINTIS ², KATARINA C. POOLE ²,
FABIAN BRINKMANN ⁴, LORENZO PICINALI ², AND MICHELE GERONAZZO ^{2,5} (Senior Member, IEEE)

¹Centre for Digital Music, Queen Mary University of London, E1 4NS London, U.K.

²Audio Experience Design, Dyson School of Design Engineering, Imperial College London, SW7 2DB London, U.K.

³Department of Neuroscience, Biomedicine and Movement, University of Verona, 37134 Verona, Italy

⁴Audio Communication Group, Technische Universität Berlin, 10587 Berlin, Germany

⁵Department of Engineering and Management, University of Padova, 35122 Padova, Italy

CORRESPONDING AUTHOR: AIDAN O. T. HOGG (e-mail: a.hogg@qmul.ac.uk).

This work was supported in part by the IEEE Signal Processing Society and in part by the SONICOM project, funded by Horizon 2020 Research and Innovation funding programme under Grant 101017743.

ABSTRACT Head-related transfer functions (HRTFs) often play a crucial role in spatial hearing, immersive audio applications for virtual reality (VR) and augmented reality (AR), and help in improving hearing assistive devices. The Listener Acoustic Personalisation (LAP) challenge 2024 aimed at advancing research in spatial audio personalisation, with a focus on head-related transfer functions (HRTFs). The challenge was split into two tasks: Task 1 was on HRTF harmonisation, and Task 2 dealt with spatial HRTF upsampling. This paper presents the results and reports the findings related to Task 2 of the LAP challenge. The submissions to Task 2 employed both algorithmic and machine learning-based approaches, which were evaluated on three key spatial audio objective metrics, including the log-spectral distortion (LSD), interaural time difference (ITD), and the interaural level difference (ILD). The results highlighted the strengths and limitations of various upsampling techniques, with learning-based methods demonstrating superior performance at lower sparsity levels. In terms of the LSD, seven of the submissions achieved an impressive performance of less than 5 dB when upsampling from only three measurement points. The results also highlighted that most submissions were often not able to outperform a generic HRTF created by averaging the HRTFs in the training dataset. One of the main contributions of this paper is that it showcases the limitations of objective metrics when it comes to evaluating HRTF upsampling. Therefore, this paper argues that a more holistic approach is needed going forward, which should include the integration of multiple perceptually relevant measures, as this is the only way to ensure a well-rounded assessment of HRTF upsampling quality.

INDEX TERMS Computational auditory modelling, head-related transfer function, upsampling.

I. INTRODUCTION

Remote Remote interaction has grown significantly in recent years; however, many challenges remain. A common issue is the lack of immersive audio in these virtual interactions. Immersive audio is what people experience in their everyday lives; some sounds are close, some are far away, some are moving, some are static, and all come from different directions. The loss of the acoustic spatial dimension, as well as the physical interactions with sound, can

lead to the frustration people often feel when communicating remotely (e.g. [1]). This becomes immediately apparent in online meetings, where multiple participants may try to speak at once and struggle to hear one another. This problem worsens when some participants are present in person and others online. However, this need not be the case, and realistic immersive audio attempts to contribute to solving this problem by making the transition from real to virtual audio seamless.

Furthermore, improvements to online communication are not the only aim. In recent years, the prevalence of VR and AR devices, as well as three-dimensional (3D) video games, auditory displays, and hearing assistive devices [2], has led to a demand for more realistic spatial audio rendering in mixed reality (MR) [3]. Achieving immersive and realistic auditory experiences in MR environments is mainly dependent on advances in modelling HRTFs, which plays a crucial role in enhancing spatial audio technologies by modelling the acoustic contribution of the listener's head, external ear, and body that allows auditory cues for spatial sound perception [4].

The LAP Challenge¹ provided an opportunity for researchers to advance the study of HRTFs and work toward establishing standardised metrics for the evaluation of personalisation approaches. It was designed as a platform to explore some of these open challenges, advance the state-of-the-art, and contribute to the development of standardised metrics for personalised spatial audio. The first edition of the challenge ran between March and August 2024 and included two tasks focused on fundamental aspects of HRTFs. A technical report [5] was published after the challenge, outlining the rules, baseline systems, and results. This paper builds upon that report by providing a detailed analysis of all submitted methods for Task 2, introducing new evaluation metrics and statistical tests, and distilling broader lessons for future research in HRTF upsampling. The focus of Task 1 was HRTF harmonisation for the purpose of merging different HRTF datasets. At the same time, Task 2 dealt with spatial HRTF upsampling, a process that increases the spatial resolution of HRTFs by estimating missing data points to improve spatial audio accuracy. In this paper, we specifically provide the outcome and results of Task 2 and report the findings related to HRTF upsampling. The results and discussion surrounding Task 1 can be found in an accompanying paper [6].

It is worthwhile to note that the HRTF processing and personalisation methods are often based on narrow tasks, i.e., specialised for the specific application/problem. Most HRTF research primarily focuses on the direction of arrival (DOA) of a sound source [7], which is a crucial yet inherently constrained perspective compared to the broader topic of spatial perception, e.g., externalisation [8], distance [9], and plausibility [10], to name but a few. Existing HRTF models typically rely on isolated metrics such as Interaural Time Difference (ITD), Interaural Level Difference (ILD), and spectral cues [11]. While this approach benefits from a relevant body of work in this domain, it proves insufficient for the complexity and variability encountered in real-world sound localisation.

In this paper, we propose a rigorous three-step framework for an objective evaluation of HRTF models in the narrow perspective of a localisation task.

- 1) *Characterise solutions*: analyse HRTF model using spatial metric distributions to identify patterns and correlations.

- 2) *Assess robustness*: explore and evaluate the resilience of solutions to changes in the metric weighting.
- 3) *Simulations*: incorporate psychoacoustically informed weighting factors based on perceptual mechanisms.

The last step is supported by a computational auditory model that predicts static DOA localisation performances of human listeners [12]. The perceptually weighted and spatially distributed metrics assessed in the previous two steps compose the feature space of the auditory model. By embracing a multifactor strategy, the proposed evaluation framework is ready to be task-inclusive, integrating additional metrics, optimisation processes, and parallel auditory simulations to enhance adaptability and accuracy in diverse applications.

II. BACKGROUND

One way to achieve high realism in immersive audio is to place the listener in the centre of a large spherical loudspeaker array and play different sounds from different directions in space. Unfortunately, although this solution works well in an acoustic lab and could be used in cinemas and other large venues, it is not practical beyond these controlled settings. It is also costly and only works well for a small number of participants at the centre of the array. A more practical solution exploits the fact that humans have two auditory sensors (i.e. two ears); in theory, we should only need two speakers (i.e. in-ear headphones) to generate the correct sound at those sensors, performing what is commonly referred to as binaural (i.e. involving the two ears) spatialisation.

A. THE MATTER OF INDIVIDUALISATION

A central challenge in binaural spatialisation lies in accurately recreating the sounds at the entrances to the ear canals to mimic real-world 3D audio accurately [13] and, more specifically, how this can be adapted for individual listeners. This individualisation has resulted in a large amount of research focusing on HRTFs, which capture the filtering effects related to the anatomy of different listeners. This filtering is caused by the sound wave being reflected and scattered off the head, torso, and pinnae before it enters the ear canal of a given listener. HRTFs are, therefore, able to capture interaural (i.e. the difference heard between the listener's two ears) and monaural localisation cues [14]. It is customary to refer to HRTF when considering the frequency domain and head-related impulse response (HRIR) in the time domain. In this paper, we also use the term HRTFs to refer to the complete set of impulse response (IR) measurements corresponding to a full set of source positions around each listener.

It has been shown in the past that using non-individualised HRTFs can significantly affect the individual's sound source localisation accuracy [15], [16], [17] as spectral cues are highly dependent on the listener's anatomy, particularly the shape of their pinnae [18]. This non-individualisation can also affect perceptual attributes such as externalisation, immersion, colouration, realism and relief/depth [19], [20], [21]. Furthermore, the choice of HRTF can significantly impact an

¹[Online]. Available: www.sonicom.eu/lap-challenge/

individual’s ability to understand speech in a cocktail party scenario [22].

As a result, capturing personalised HRTFs remains an important area of active research. It has been shown in the past that many approaches can be deployed for this task of HRTF individualisation, including acoustic measurements [23], 3D scans [24], modelling the morphological geometric information of the listener’s ears [25], [26] and selecting the best-fitting HRTF from a database of already-measured ones. This best-fitting HRTF selection is often made using morphology-based methods [27], [28] or perceptual-based methods (e.g. using individual preference [29] and/or localisation accuracy [30], [31]). An overview of some of the most common methods can be found in [11].

The acoustic measurement [32] is still considered the gold standard of these different approaches. The downside to performing this acoustic measurement is the expensive custom setup required and the time it takes. This is because numerous IRs need to be measured around the individual, with numbers ranging anywhere from 200 to 3000 [23]. This process can be sped up by taking advantage of interlaced sine sweeps [33], but this often only makes the elevation measurements faster. Other methods do exist [34], [35] that aim to improve the time performance of the HRTF measurement, but the equipment specifications and cost are usually very high.

B. SPATIAL UPSAMPLING OF HRTFS

To reduce the time required and the complexity of the HRTF setup and to make the method scalable, spatial upsampling methods have been proposed in the past that can generate high-resolution HRTFs, i.e. HRTFs that contain many (normally over 300) IRs from many directions, from low-resolution HRTFs, i.e. HRTFs that include very few IRs from very few directions [36]. This process is commonly referred to as HRTF upsampling and can be achieved using various approaches.

Two fundamentally different HRTF upsampling approaches can be distinguished: algorithmic and learning-based approaches. Algorithmic approaches usually *interpolate* HRTFs for new source positions by a superposition of HRTFs or basis functions generated thereof. The interpolation is often done independently for each frequency or time sample using only the sparse HRTF data set or even a sub-set thereof as input (cf. Fig. 1 top). In contrast to this, learning-based approaches usually *generate* HRTFs for new source positions using neural networks that are trained on data sets of high spatial resolution under the premise that the upsampling must later work based only on spatially sparse data (cf. Fig. 1, bottom). This opens the potential to learn relations between HRTFs for different source positions, frequencies and time samples. This is an advantage over algorithmic approaches, where such knowledge can only be considered to a limited extent, e.g., by modelling the time-of-arrival or considering magnitude spectra in auditory resolution (for example, see [37]). On the other hand, algorithmic approaches work on (almost) arbitrary sets of source positions, whereas neural

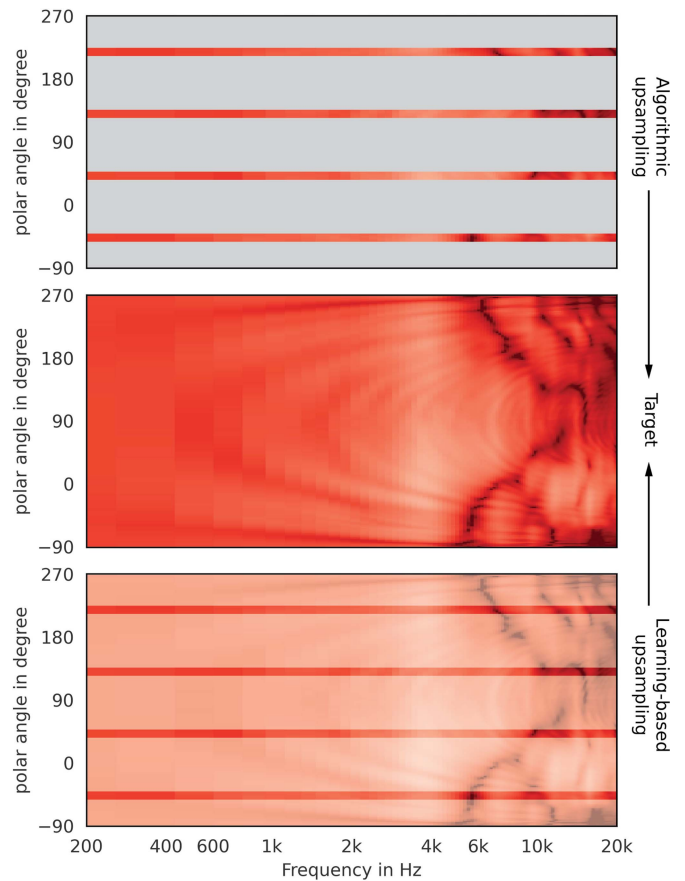


FIGURE 1. Schematic illustration of algorithmic and learning-based upsampling in the median plane using four HRTFs. The top and bottom images illustrate the sparse input HRTFs that are available for algorithmic (top) and learning-based (bottom) upsampling. The target HRTFs are shown in the middle. Grey rectangles denote data that is unavailable for algorithmic upsampling. In the case of learning-based upsampling, opaque data is only available during training but not during inference. Colour codes the HRTF magnitude between -20 dB (dark red) and 20 dB (light red).

networks are trained for a specific set of source positions. Additionally, the latter may generate erroneous HRTFs due to hallucination or similar errors (for example, see Fig. 3, bottom-right), while algorithmic approaches are robust to unseen input data—at least if well parametrised.

The most common algorithmic HRTF upsampling method is barycentric interpolation [38], [39], [40]. This method has been shown to produce good results when the HRTFs contains a relatively large number of IRs [41], for example, with an angular distance of 10-15° between measurements; however, it becomes much less reliable when interpolating sparser measurements (e.g. each 30-40°). Another common approach is spherical harmonic (SH) interpolation [37], [42], [43], [44], [45], but again results in poor reconstruction when the low-resolution HRTF input is spatially sparse. This is because these methods rely on averaging between existing data points based on prior information. For example, barycentric interpolation uses the three nearest neighbours around the point to be interpolated to calculate the weighted average. Therefore, as

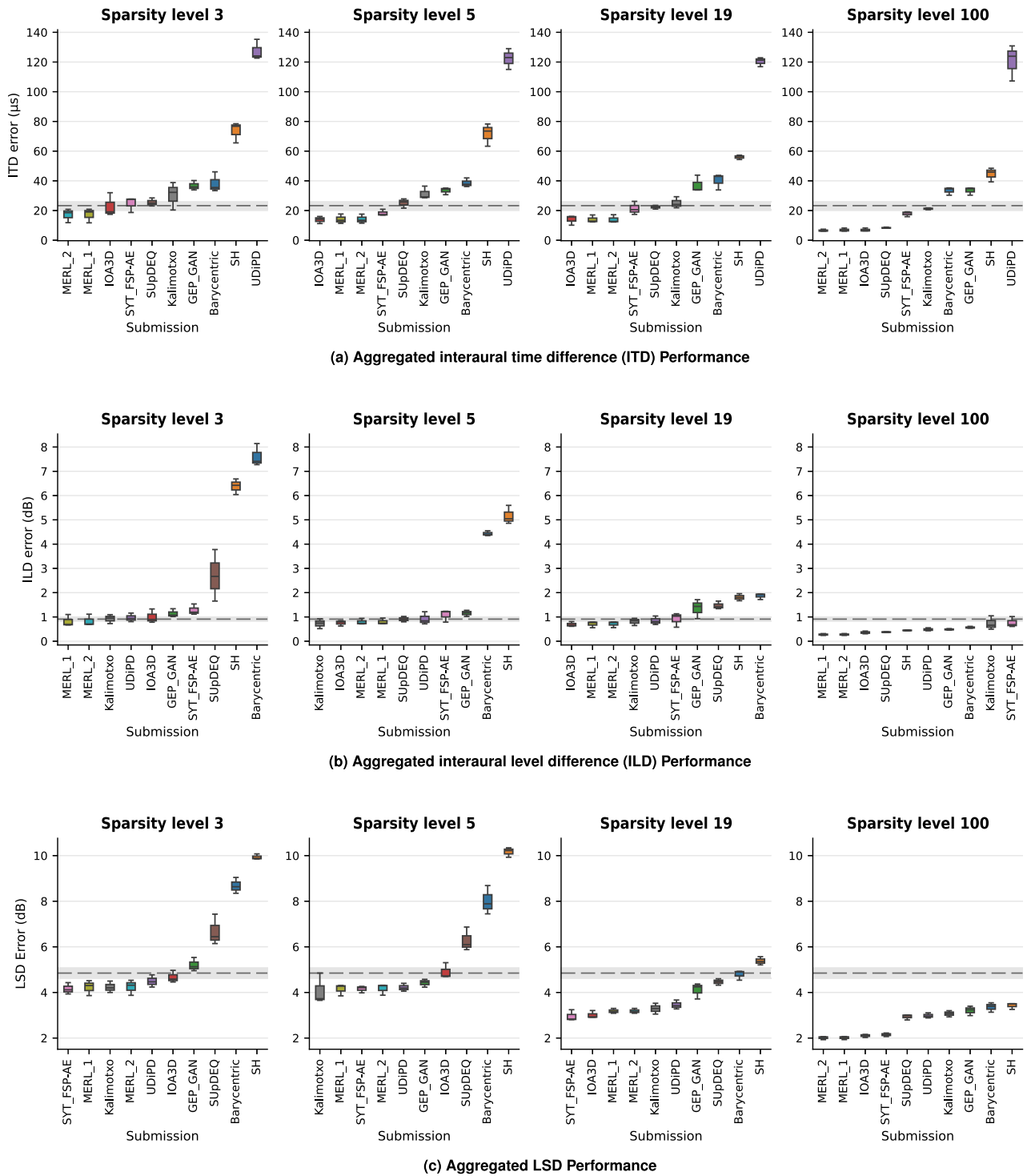


FIGURE 2. Aggregated performance across all sparsity levels for ITD, ILD, and LSD. The dashed line and shaded areas represent the mean and standard deviation of the error metrics calculated between the original HRTF and an averaged HRTF created from the HRTFs in the SONICOM dataset.

the distances between the neighbours grow larger, the upsampling becomes more and more inaccurate.

In the last 5 years, machine learning (ML) methods have started to become the focus of research on HRTF personalisation [46]. In the past, ML techniques have been shown to be effective at estimating HRTFs from just the anthropometric

measurements of the listener. In [47], a deep neural network (DNN)-based approach is used to synthesise a personalised HRTF using the anthropometric features of the user and was able to achieve a performance of 3.2 dB LSD. This approach consisted of using the encoder part of an autoencoder to reduce the dimensionality of the raw HRTFs, which can be used

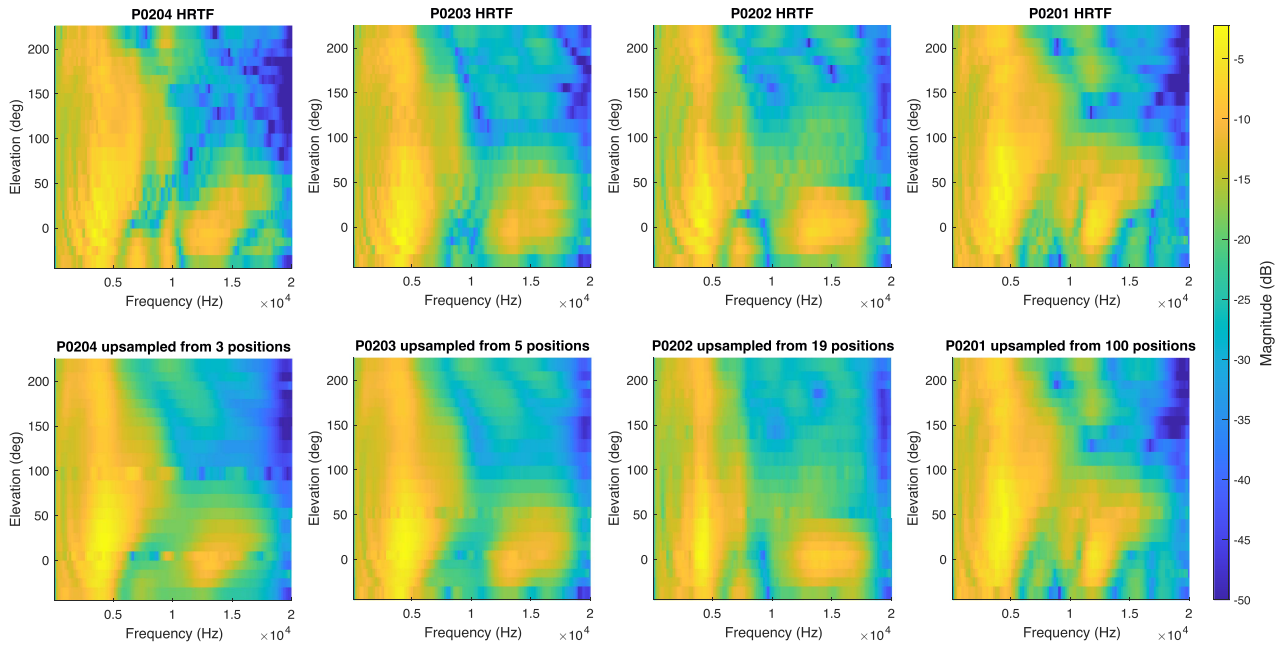


FIGURE 3. Median plane spectra of example upsampled HRTFs using the winning solution (MERL 2) compared to the original HRTFs for each upsampling factor.

as a set of input features for training. This aims to minimise overfitting as HRTF datasets are usually small. The decoder part of the autoencoder then estimates the HRTF magnitudes using the output of a DNN that is trained to output the latent representation given the anthropometric features and the target azimuth. This type of approach is explored further in [48], where two autoencoders are exploited: one that reduces the dimensionality of a feature vector containing the azimuth and anthropometric parameters and another that reduces the dimensionality of the magnitudes of the full HRTF measurement. These two autoencoders are combined to estimate the HRTF magnitudes from anthropometric features and achieve a performance of 4.3 dB LSD.

It has also been shown in the past that autoencoders can be used to upsample low-resolution HRTFs. In [49], a method is proposed that uses an autoencoder and is an extension of a regularised linear regression (RLR) approach that makes use of the spherical wavefunction presented in [50]. This method's key feature is that it decomposes HRTFs into source position-dependent and source position-independent factors, i.e. the spherical wavefunction expansion and expansion coefficients, respectively. The autoencoder is conditioned on source positions and obtains the source-position-independent representation by using an aggregation module between the encoder and decoder, aggregating latent variables of a given source position. This approach was able to notably achieve 4.4 dB in LSD when upsampling from 9 to 440 positions. Other ML methods also exist that are able to perform the task of HRTF upsampling, including [51], which exploits a deep belief network (DBN). This method accomplishes an LSD average of less than 3 dB; however, results are only given for upsampling from 125 source positions to 1250,

which is still relatively dense. Another method in [52] uses a convolutional neural network (CNN) and has been shown to yield a good performance of 4.4 dB LSD when upsampling from 23 positions to 1250 and 3.8 dB when upsampling from 105 positions. However, in this case, the sphere is sliced into planes to create a two-dimensional (2D) representation rather than considering the sphere as a whole. ML techniques have also been used in the past as part of a post-processing step of the spherical harmonic transform (SHT) interpolation [53].

In [54], [55], another novel method is proposed that combines both algorithmic and learning-based approaches to HRTF spatial upsampling, employing an autoencoder generative adversarial network (AE-GAN) architecture in the SH domain. By transforming frequency-domain HRTFs into SH coefficients, their model achieves a compact and structured representation that enables effective extrapolation even under extreme spatial sparsity, fewer than 12 measurements. The evaluations, both perceptual and spectral, suggest that the SH coefficient extrapolation process naturally aligns with sequence-to-sequence prediction tasks, opening the door for future exploration with recurrent neural networks (RNNs) and transformers to enhance upsampling accuracy. More recently, this work has utilised U-Net models due to their ability to generate high-fidelity, individualised data from not only sparse but also noisy measurements [56].

Diffusion-based models also present a promising avenue for future research in HRTF personalisation [57]. These models offer the potential to capture the complex, high-dimensional structure of HRTFs and generate perceptually plausible results even in extremely sparse conditions. Integrating such models with perceptually grounded evaluation frameworks, as

discussed later in this study, may also help mitigate some of the limitations that currently exist with deep learning methods.

III. TASK DEFINITION

This paper focuses on Task 2 of the LAP challenge, which involved reconstructing high-resolution HRTFs from sparse measurements. The SONICOM HRTF dataset [23], a publicly released resource designed to support reproducible research in spatial acoustics and immersive audio, was used in this challenge. Each HRTF in the dataset consists of impulse responses measured at 793 spatial locations, with azimuth angles ranging from 0° to 355° (in 5° steps) and elevation angles from -45° to 90°.

Two distinct subsets of the SONICOM dataset were used. The first subset consisted of 200 publicly available HRTF sets, which were made available to participants for training. These 200 HRTFs were also averaged to provide a reference for the task evaluation in Section VI. Separately, a dataset of 12 sparse HRTFs, derived from previously unreleased subjects, was used in the design of the task. Each sparse HRTF represented measurements from a unique individual and was provided at one of four sparsity levels (three subjects per level, totalling twelve). Participants were asked to reconstruct dense HRTFs from these sparse inputs. The corresponding ground-truth dense versions were withheld so submissions could be evaluated against these to assess reconstruction quality.

The following definitions characterise the levels.

- *Low up-sampling factor*: 100 positions around the listener. The positions are selected by sorting 793 SONICOM HRTF positions by azimuth (from 0° to 355°) and elevation (from -45° to 90°) and keeping every 8th position, starting from the 1st (at 0° azimuth -45° elevation).
- *Mid-up-sampling factor*: 19 positions around the listener. Six azimuth positions with 60° spacing (0° to 300°) and three elevations (-45°, 0°, and 45°) are selected, plus the top position (0° azimuth, 90° elevation).
- *High up-sampling factor, non-uniformly spaced*: 5 positions clustered around one direction. Positions are centred within 45° around the front of the listener (0° azimuth and ± 45° elevation, ± 45° azimuth and 0° elevation, and 0° azimuth/elevation).
- *Very high up-sampling factor*: 3 positions. One is in front of the listener (0° azimuth and elevation), one to the left (90° azimuth and 0° elevation), and one above (0° azimuth and 90° elevation).

Table 1 provides information on the specific HRTFs used in the challenge. Four sparsity levels (100, 19, 5, and 3 positions) were used, with 3 HRTFs provided for each level. The sparse HRTFs are obtained from the versions of the full SONICOM HRTFs that have been compensated using a minimum phase-free field compensation filter and are sampled at 48 kHz [23]. Original versions of these HRTFs were released after the submission deadline for the challenge.²

²[Online]. Available: https://transfer.ic.ac.uk:9090/#/2022_SONICOM-HRTF-DATASET/

TABLE 1. HRTFs and Their Sparse Versions Used for Task 2

Original HRTF	Sparse HRTF
P0201_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_100_1.sofa
P0205_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_100_2.sofa
P0210_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_100_3.sofa
P0202_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_19_1.sofa
P0206_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_19_2.sofa
P0211_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_19_3.sofa
P0203_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_5_1.sofa
P0207_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_5_2.sofa
P0212_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_5_3.sofa
P0204_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_3_1.sofa
P0208_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_3_2.sofa
P0213_FreeFieldCompMinPhase_48kHz.sofa	LAPtask2_3_3.sofa

The details provided to the team participating in the challenge are available online.³

IV. TASK EVALUATION

The evaluation focused on the similarity of upsampled HRTF and their original versions [58]. This evaluation was conducted using Python scripts, which determine the differences in spatial feature values of the upsampled HRTFs against the original datasets and are available at the official GitHub repositories⁴ and.⁵

To objectively assess the performance of HRTF upsampling methods, three key perceptually motivated metrics are used for the evaluation: LSD, ITD, and ILD. These metrics evaluate the spectral, temporal, and level-related fidelity of upsampled HRTFs, respectively.

These metrics aim to evaluate the spectral, temporal, and level-related fidelity of upsampled HRTFs, respectively.

- *Log-Spectral Distortion (LSD)* quantifies the spectral similarity between a reference HRTF and an upsampled estimate across directions and frequency bins. Following [59], it is defined as:

$$\text{LSD} = \frac{1}{N} \sum_{n=1}^N \sqrt{\frac{1}{B} \sum_{b=1}^B \left(20 \log_{10} \frac{|H_{\text{HR}}(f_b, x_n)|}{|H_{\text{US}}(f_b, x_n)|} \right)^2}, \quad (1)$$

where $|H_{\text{HR}}(f_b, x_n)|$ and $|H_{\text{US}}(f_b, x_n)|$ denote the magnitude spectra of the high-resolution and upsampled HRTFs at frequency bin f_b and spatial location x_n , B is the total number of frequency bins, and N is the number of spatial directions. The LSD is reported in decibels (dB) and captures spectral discrepancies affecting timbral perception.

³[Online]. Available: <https://www.sonicom.eu/lap-challenge/>

⁴[Online]. Available: <https://github.com/Audio-Experience-Design/LAPChallenge>

⁵[Online]. Available: <https://spatial-audio-metrics.readthedocs.io/en/latest/>

- *Interaural Time Difference (ITD)* quantifies the error in ITD estimation between the upsampled and high-resolution HRTFs. The ITD error is computed as the average absolute difference across all spatial locations:

$$\text{ITD} = \frac{1}{N} \sum_{n=1}^N \left| t_{\text{HR}}^{\text{Left}}(x_n) - t_{\text{HR}}^{\text{Right}}(x_n) \right| - \left| t_{\text{US}}^{\text{Left}}(x_n) - t_{\text{US}}^{\text{Right}}(x_n) \right|, \quad (2)$$

where $t_{\text{HR}}^{\text{Left}}(x_n)$ and $t_{\text{HR}}^{\text{Right}}(x_n)$ are the time-of-arrival estimates for the left and right ears at location x_n from the high-resolution HRTF, and $t_{\text{US}}^{\text{Left}}(x_n)$ and $t_{\text{US}}^{\text{Right}}(x_n)$ are the corresponding values from the upsampled HRTF. The ITD error is the average absolute difference between high-resolution and upsampled ITDs and is expressed in microseconds (μs). It reflects timing discrepancies in localisation.

- *Interaural Level Difference (ILD)* reflects the level difference between ears across frequencies and spatial locations. Based on [45], [60], it is defined as:

$$\text{ILD} = \frac{1}{N} \sum_{n=1}^N \frac{1}{B} \sum_{b=1}^B \left| \left(20 \log_{10} \frac{|H_{\text{HR}}^{\text{Left}}(f_b, x_n)|}{|H_{\text{HR}}^{\text{Right}}(f_b, x_n)|} \right) - \left(20 \log_{10} \frac{|H_{\text{US}}^{\text{Left}}(f_b, x_n)|}{|H_{\text{US}}^{\text{Right}}(f_b, x_n)|} \right) \right|, \quad (3)$$

where $|H^{\text{Left}}(f_b, x_n)|$ and $|H^{\text{Right}}(f_b, x_n)|$ represent the magnitude spectra for the left and right ears, respectively. The ILD is expressed in decibels (dB), and discrepancies may result in lateral perceptual shifts or degraded spatial clarity.

The upsampled HRTFs are compared against ground-truth measurements using the three metrics. Each metric is computed over the full spherical grid and then averaged per subject and condition. Metrics are further normalised and aggregated to produce final rankings.

These metrics provide a perceptually grounded basis for assessing HRTF upsampling quality. While they offer quantitative insight, their limitations, particularly regarding perceptual relevance, are further addressed in later sections via auditory model-based evaluation and discussion.

The challenge uses these metrics as part of a two-stage evaluation.

Stage 1 - The upsampled HRTFs are validated against threshold values for the ITD, ILD, and LSD. These thresholds were derived from baseline interpolation methods (barycentric and spherical harmonic) and were set to:

- *ITD*: 100 μs ,
- *ILD*: 4.4 dB,
- *LSD*: 7.4 dB.

To advance to the next stage, submissions had to meet the thresholds for at least 10 out of 12 HRTFs.

Stage 2 - Performance was evaluated by calculating the average metrics across all upsampling factors. The results were then normalised, and the final ranking was determined based on the lowest total z-score across all sparsity levels.

V. OVERVIEW OF SUBMISSION METHODOLOGIES

The 2024 edition of the LAP challenge received eight submissions, and here follows a summary of the submitted methods:

- *UDiPD*: the authors improve a method that combines neural field-based spatial upsampling with differentiable IIR filters, whose parameters are optimised via back-propagation to minimise binaural filter response errors through a perceptually motivated loss function.
- *IOA3D*: this approach proposes a convolutional neural network-based model with an upsampling-downsampling architecture, which consists of three components: (i) an encoder extracting the extrapolation features from sparse measurements, (ii) a decoder reconstructing input measurements, and (iii) a separate network predicting high-resolution HRTFs.
- *SUPDeq*: the solution by Pörschmann, Brinkmann, and Arend is based on the Spatial Upsampling by Directional Equalisation (SUPDeq) method using natural-neighbour interpolation of Magnitude-Corrected and Time-Aligned (MCA) HRTFs and an analytical spherical head equalisation function [37]. For sparsity levels 3 and 5, magnitude correction was omitted in favour of a custom equalisation function combining the phase of the spherical head model with the one-six octave smoothed and averaged magnitude response of the first 200 SONICOM HRTF sets. For those sparsity levels, the HRTFs were mirrored from the left to the right and from the front to the back prior to the interpolation, but after the directional equalisation. This is the only algorithmic approach and was submitted to provide a sensible baseline for the learning-based approaches.
- *GEP_GAN*: the method leverages existing image-upscaling literature, which converts spherical HRIR data into a format for convolutional neural network processing using two steps: projection onto a 2D surface and interpolation onto an evenly spaced Cartesian grid.
- *MERL_1 and MERL_2*: the proposed method uses a neural field for personalised HRTF prediction, parameter-efficient fine-tuning to efficiently adapt the model with few samples per subject, and retrieval augmented generation to retrieve similar subjects' data.
- *SYT-FSP-AE*: the proposed method is an extension of a previous upsampling method for HRTF magnitudes using a conditioned autoencoder that takes source position and frequency as input, allowing the use of diverse training data from different spatial grids.
- *Kalimotox*: the method uses an autoencoder architecture with a neural network to upsample and reconstruct

TABLE 2. The Z-Score Ranking Summary for the Seven Task 2 Submissions That Passed the Stage 1 Validation Thresholds

Overall		Sparsity level 3		Sparsity level 5		Sparsity level 19		Sparsity level 100	
Model	Z-Score	Model	Z-Score	Model	Z-Score	Model	Z-Score	Model	Z-Score
MERL_2	-58.098	MERL_1	-12.688	MERL_1	-13.329	IOA3D	-14.732	MERL_2	-17.600
MERL_1	-58.075	MERL_2	-12.674	MERL_2	-13.312	MERL_1	-14.514	MERL_1	-17.543
IOA3D	-55.688	IOA3D	-11.002	IOA3D	-12.612	MERL_2	-14.510	IOA3D	-17.341
SYT_FSP-AE	-51.355	SYT_FSP-AE	-10.922	SYT_FSP-AE	-12.270	SYT_FSP-AE	-13.241	SUPDEQ	-16.165
Kalimotxo	-46.457	Kalimotxo	-10.249	Kalimotxo	-10.488	Kalimotxo	-12.353	SYT_FSP-AE	-14.920
SUPDEQ	-42.256	GEP_GAN	-7.848	GEP_GAN	-9.264	SUPDEQ	-10.773	Kalimotxo	-13.366
GEP_GAN	-37.182	SUPDEQ	-6.360	SUPDEQ	-8.956	GEP_GAN	-8.696	GEP_GAN	-11.372

HRTF magnitudes from specific elevations based on original measurements, previous reconstructions, and average HRTFs for each direction.

VI. SUBMISSION RESULT ANALYSIS

Out of the submitted upsampled HRTFs, the ones from **UDiPD** did not pass the validation step with 12 out of 12 invalid HRTF datasets due to large ITD errors, and the validation failed for the **SUPDEQ** submission with the HRTF dataset named `LAPTask2_3_1.sofa`. Seven out of eight submissions that passed the validation step moved on to the second stage of evaluation, where their performance was aggregated across the different upsampling factors. The Fig. 2(a), (b) and (c) provide visual representations of the performance across the three metrics for all teams. Each figure reports each metric aggregated across all upsampling factors. For comparison, we also show the performance of the SONICOM average HRTF. It was generated by separately averaging the magnitude responses and times of arrival of the first 200 SONICOM subjects per source position. The times of arrival were estimated using the ten times upsampled and low-passed HRIRs (10th order Butterworth, cut-off frequency 3 kHz) by finding the first sample whose value is 10 dB below the absolute maximum of each left and right ear HRIR. In a final step, the average HRTF was generated by applying fractional delays according to the averaged times of arrival to minimum phase versions of the averaged HRTF magnitude spectra.

The final ranking was based on the combined performance across all sparsity levels since the overall performance of the teams was assessed by summing their z-scores across all upsampling factors. This combined score determined the final ranking of the submissions. The results are summarised in Table 2. The winner for this task is the submission named **MERL_2**. Finally, Fig. 3 illustrates the spectral cues over the median plane (i.e. azimuth set to zero) of the HRTF upsampled using the winning solution.

The results presented in Fig. 2 clearly only tell part of the story. It is, therefore, essential to develop a deeper understanding of what these results signify and to evaluate whether the challenge successfully assessed HRTF upsampling effectiveness. Additionally, Fig. 3 reveals that the winning solution, **MERL_2**, still exhibits significant differences from the original

HRTF. However, determining where these differences matter perceptually remains an open challenge.

A. ANALYSING THE LSD SPATIAL DISTRIBUTIONS

One limitation of the evaluation framework is that the metrics are averaged over the entire sphere. However, perceptual studies indicate that certain spatial regions are more critical for auditory perception than others. Fig. 4 visualises the spatial distribution of errors, showing that had less emphasis been placed on elevation -50° (a region where localisation is perceptually less relevant), the SUPDEQ and GEP-GAN methods would have performed significantly better at the lowest sparsity level, as both perform poorly in this region. A positive observation from Fig. 4 is that both algorithmic and learning-based approaches exhibit similar error distributions at higher upsampling factors, suggesting a level of consistency across approaches.

B. ASSESSING THE ROBUSTNESS OF THE EVALUATION

It is also important to assess the robustness of the evaluation framework, particularly regarding the challenge winner. When designing the LAP challenge evaluation metric, several key decisions were made, including the weighting of the LSD, ITD, and ILD objective metrics. Now that the challenge has concluded, it is important to assess how these weighting choices may have influenced the final rankings. Figs. 5 and 6 illustrate how metric weighting impacts the rankings of different methods, including the winning solution. Notably, **MERL_2** demonstrates strong robustness to changes in metric weighting, reinforcing its reliability. In contrast, Fig. 6 highlights a notable example: the SYT-FSP-A method undergoes substantial ranking shifts depending on the chosen metric. Specifically, in Fig. 6(a), this method ranks first when only LSD is considered, seventh when only ILD is used, and fourth when ITD is the sole metric. This raises important questions about the relative perceptual significance of each metric and whether equal weighting is a sensible approach.

C. ANALYSIS OF HRTFS

In Fig. 7, examples of upsampled HRTFs are given for two sparsity levels and a source located to the right. This example for sparsity level 3 is among the most challenging source positions for upsampling within the LAP challenge because

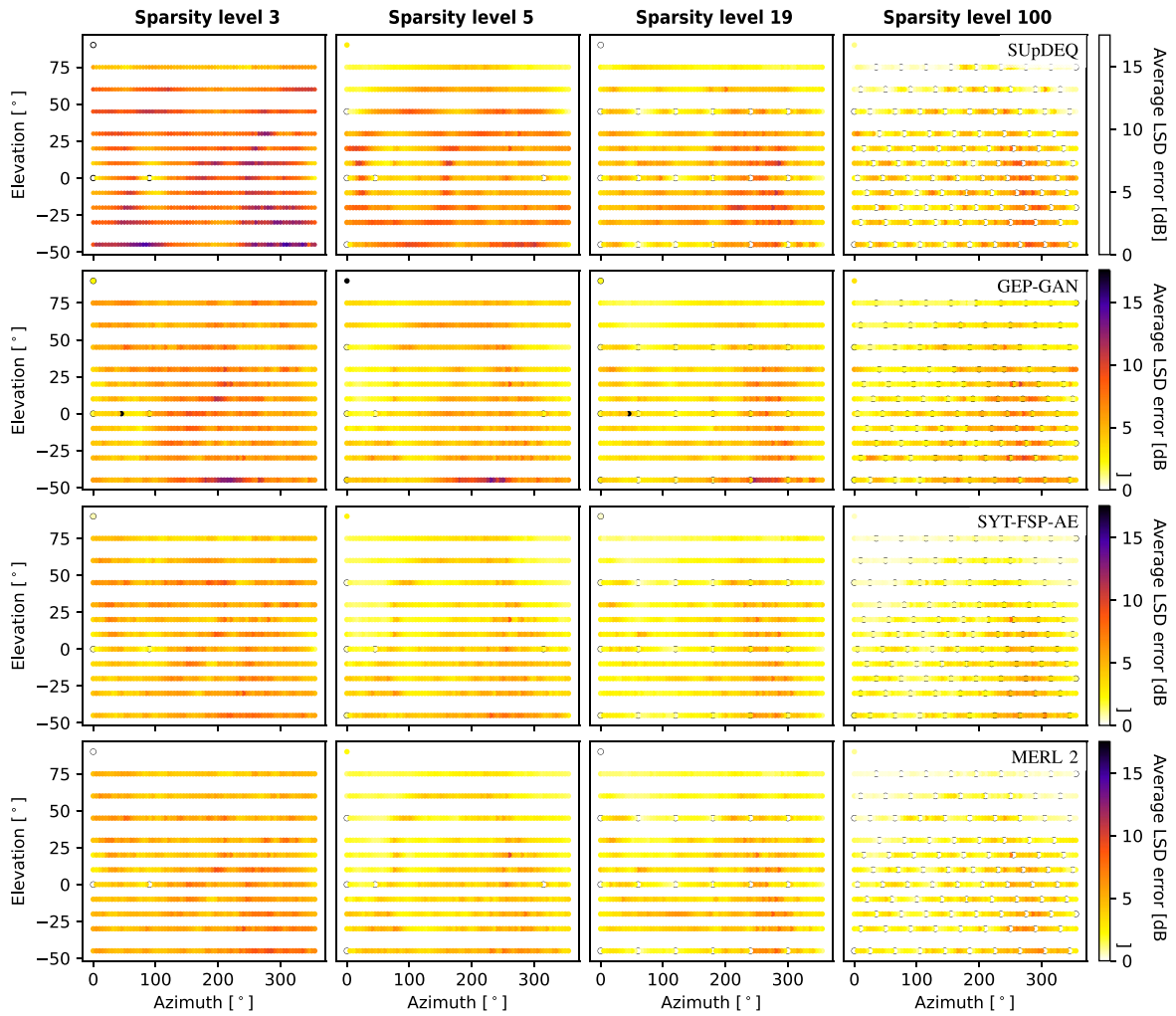


FIGURE 4. LSD distribution for Subject 1 across different HRTF upsampling methods and sparsity levels. The circles indicate the original sample points provided at each sparsity level, while white areas indicate no sample points at that elevation. Top row: SUPDEQ-MCA (algorithmic-based), second row: GEP-GAN (algorithmic and learning-based), third row: SYT-FSP-AE (learning-based), bottom row: MERL 2 (learning-based, winner).

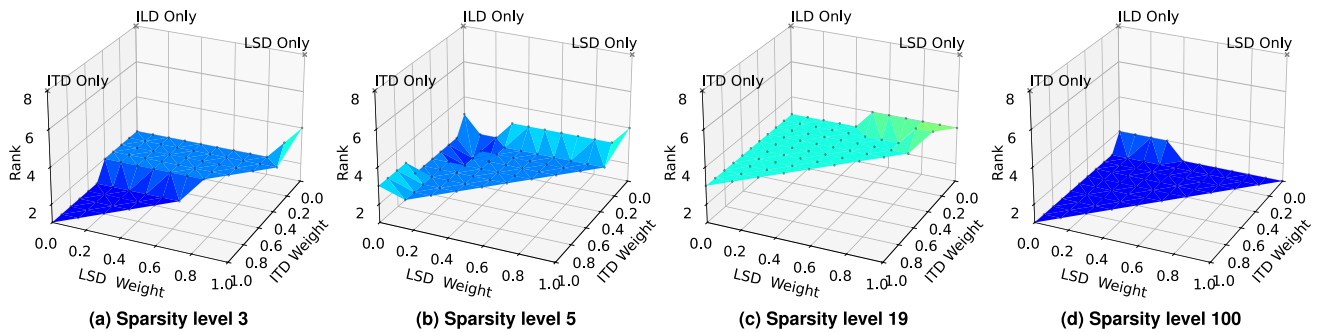


FIGURE 5. MERL 2 (winner) Ranking. This winning solution is very resilient to changes in metric weighting.

the target source has a distance of 90° to the closest HRTF contained in the sparse data set, which is located at the front (0° azimuth, 0° elevation). In this case, all upsampled HRTFs seem to be closer to the average HRTF than to the reference, regardless of the ear. For the contralateral left ear, the average HRTF is approaching the reference HRTF quite well, apart

from frequencies where spectral notches occur. However, these notches might be perceptually irrelevant. Considering that the level of the left ear is approximately 10 dB to 30 dB below that of the right ear for this source position, the right ear might be dominating the perceptual impression. This is also true for the contralateral ear, even for difficulty 100. In

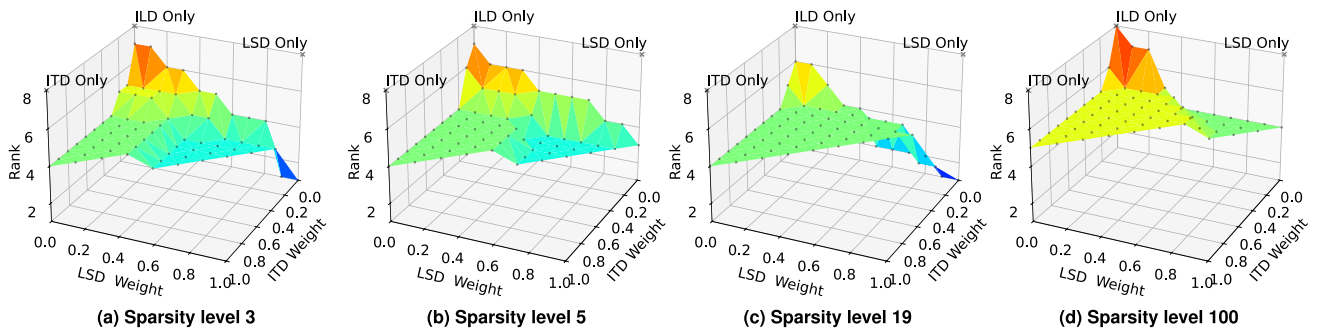


FIGURE 6. SYT-FSP-AE Ranking. This is an example where a solution is not very resilient to changes in metric weighting.

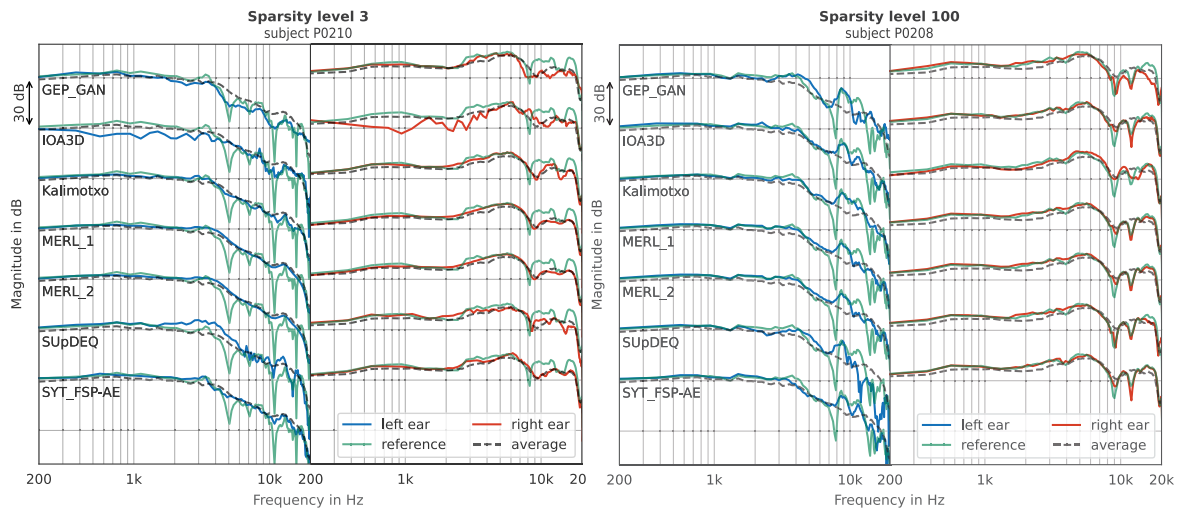


FIGURE 7. Selected upsampled HRTFs for all submissions, two sparsity levels, and a source to the right (265° azimuth, 0° elevation). The reference and average HRTF are shown for comparison. The average HRTF was computed using the magnitude responses of the first 200 subjects of the SONICOM database.

this case, however, the upsampled HRTFs are all closer to the reference than to the average HRTF for the ipsilateral right ear.

For difficulty 3, the HRTFs upsampled by IOA3D show clear artefacts for both ears. In the case of the ipsilateral ear, errors of up to 10 dB appear for frequencies between 1 kHz and 5 kHz, where the human auditory system is most sensitive. Similar artefacts were observed for other algorithms under different conditions, but they were most frequent and severe for IOA3D. Plots for all difficulty levels and subjects, as well as auralizations of moving sources in the horizontal and median plane, are contained in the supplemental material.

D. AUDITORY MODEL EVALUATION

Assessing HRTF upsampling solely through acoustic cues overlooks non-acoustic perceptual factors in spatial hearing [61]. Instead, controlled localisation experiments would provide direct perceptual validation but require extensive data collection beyond this study's scope. As an initial step, we employed an auditory model to simulate perceptual evaluation, examining how distortions in upsampled auditory cues affect

localisation performance [12]. Although other auditory localisation models exist (see [62] for a review), we selected the model by Barumerli et al. [12] because it uniquely integrates all relevant auditory features to jointly estimate localisation in both horizontal and vertical directions, includes parameters tuned to match the localisation uncertainty of listeners, is fully interpretable, and was publicly available at the time of writing, aligning well with the scope of this study.

The selected model simulates human spatial hearing in static scenarios through Bayesian inference (i.e. consistent with the acoustically measured HRTFs) and produces behavioural metrics comparable to those observed in human listeners [63]. It processes binaural inputs (i.e., the upsampled HRIRs) and outputs directional estimates, specifically lateral and polar angles. The model integrates three spatial cues (i.e. broadband ITD, ILD, and spectral cues similar to those utilised in Section IV) and combines them through an uncertainty-weighted mechanism to reflect variability in cue reliability. In particular, weights are cue-specific and informed by psychoacoustic thresholds and individual localisation performances [12]. This process involves comparing the input features with precomputed templates derived

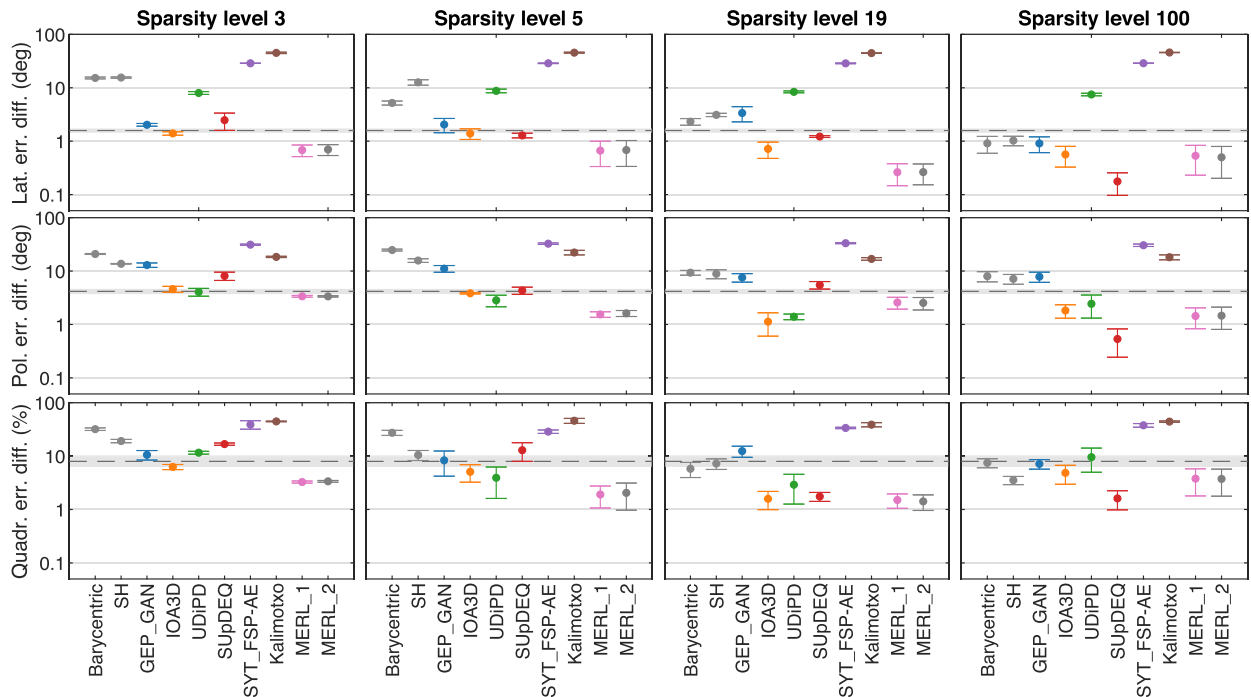


FIGURE 8. Increase in model-based localisation errors caused by upsampling across submitted upsampling methods and baseline methods (barycentric and spherical harmonic interpolation). Metrics include lateral RMS error (top), polar RMS error (middle), and quadrant error (bottom). Results are reported as absolute differences between predicted metrics with original and upsampled HRTFs. The y-axis is presented on a log scale to better compare the methods. Results are averaged over three datasets for each upsampling factor (3, 5, 19 and 100 positions). Error bars indicate the standard error across datasets. The dashed line and shaded areas represent the mean and standard error of the metric difference between the original HRTF and the SONICOM averaged HRTF datasets.

from the original HRTF dataset, each associated with specific directions, to generate a likelihood function. Then, the model integrates priors to account for perceptual biases beyond acoustic cues, such as prior experience in spatial localisation. The posterior distribution then informs the model’s output, selecting the most likely direction.

We analysed key behavioural metrics to evaluate the model’s directional estimates generated with upsampled HRTF datasets. Specifically, we evaluated lateral and polar root mean square (RMS) errors and quadrant errors [64], quantifying estimation errors in horizontal and vertical dimensions. To ensure alignment with perceptual studies, errors were examined within spherical sectors most sensitive to feature distortions: vertical errors were measured within the median plane ($\pm 30^\circ$ lateral angle). In contrast, horizontal errors were assessed in the frontal region ($\pm 60^\circ$ from the median plane). Model parameters were configured according to [63] except that the pointing error was set to zero, and, to ensure statistical robustness, each direction was estimated 300 times.

To assess how upsampling alters spatial cues, we computed localisation error differences by subtracting the simulated responses of the original HRTFs from their upsampled counterparts. Results were averaged over three datasets per upsampling factor to ensure comparability and reduce biases from sparsity levels. Additionally, we evaluated these differences against both the upsampling baselines (i.e. barycentric and

spherical harmonics interpolation) and the averaged HRTF dataset introduced in Section VI. This allowed us to examine whether upsampling methods preserved individual spatial characteristics or converged toward a generalised representation. By comparing localisation errors across these conditions, we conducted an initial model-based analysis to estimate how upsampled HRTFs retained perceptual distinctions and improved upon baseline interpolation methods.

The model-based evaluation revealed notable differences among submissions regarding their consistency with behavioural metrics (see Fig. 8). In agreement with the challenge evaluation method, the MERL submissions demonstrated the most consistent performance across both the median plane and horizontal-frontal sectors, indicating robustness in the proposed method at both global and local spatial scales. Similarly, several other methods exhibited reasonable consistency (e.g. GEP-GAN, IOA3D and SUPDEQ), suggesting that their optimisation strategies effectively balanced acoustic accuracy and localisation relevance. Notably, the approach proposed in SUPDEQ performed well at higher sparsity levels (e.g. 100 sample points) even when compared to MERL_1 and MERL_2, benefiting from its perceptually informed methods. However, deep learning-based solutions excelled in generalisation at lower sparsity levels (e.g. 3 sample points), proving effective for complex spatial auditory reconstruction. Nevertheless, MERL’s subject-dependent fine-tuning might have inflated variability in the predicted metrics across listeners for

simpler conditions, as reflected by the larger standard errors reported in Fig. 8. As a last remark, the SYT-FSP-AE and Kalimotxo submissions exhibited higher errors than baseline methods, indicating that global error optimisation might fail to preserve local acoustic cue coherence. In contrast, UDiPD achieved low polar and quadrant errors but incurred high lateral error, indicating difficulty in accurately reconstructing ITDs as reported in Fig. 2(a).

We also compared localisation errors between upsampled and averaged HRTFs to evaluate whether participants' methods preserved individual spatial cues or converged to a generalised HRTF dataset. The results revealed that, for some methods, as the number of source positions decreased, upsampling errors increased, eventually surpassing the localisation differences between individual and averaged HRTFs. This suggests that certain upsampling methods might approximate a generalised HRTF in highly sparse conditions rather than preserving listener-specific spatial cues. Additionally, the evaluation exposed an interesting phenomenon: while spectral averaging introduced in the averaged HRTF dataset removes fine-grained local features by levelling notches and peaks, which are critical for vertical localisation [65], some spatial information remains accessible in the model evaluation. This may explain why specific methods performed well when evaluated over the full sphere or large spherical sectors despite failing to preserve localised spatial features. The evaluation methodology is a potential reason for this: model-based and acoustic cue assessments rely on frequency-domain averaging without perceptual weighting, potentially overlooking distortions in psychoacoustically critical bands [66]. These findings highlight the risk of upsampling and reducing spatial personalisation, as a shift toward an averaged HRTF may compromise perceptual accuracy despite its usefulness in some applications, making perceptual validation essential to determine whether listeners perceive these differences.

VII. DISCUSSION AND LIMITATIONS

The results of the LAP challenge provide valuable insights into HRTF upsampling performance; however, several key factors need to be carefully considered to further refine the evaluation process and improve future iterations of the challenge.

One notable observation is that Task 2 was more effective than Task 1, likely due to the availability of a ground-truth reference (the measured HRTF), making it easier to validate and compare different methods. Nevertheless, while objective metrics provide a helpful baseline, perceptual evaluation remains essential for a comprehensive assessment of upsampling quality. Future work should incorporate listening tests, combining both qualitative measures (e.g. externalisation, timbre, and preference) and quantitative assessments (e.g. localisation accuracy and spatial release from masking). Although perceptual experiments are time-consuming, a study involving 12 participants with their own measured HRTFs

upsampled by different methods could be feasible, requiring approximately two hours per subject.

Beyond evaluating current approaches, the next step in HRTF upsampling research could involve real-time, on-the-fly personalisation within AR environments. A key question remains whether upsampling a sparse set of HRTFs is superior to selecting a high-resolution, non-individualised HRTF from a database. Future studies should investigate this trade-off.

A. COMPARISON TO AVERAGE HRTF

Fig. 9 indicates in which cases the submissions perform better than the SOMICOM average HRTF. The motivation behind this comparison is that an individualisation approach could be considered useful if it outperforms a generic HRTFs, which in this case is represented by the average HRTF. This simplified and condensed overview reveals the following: MERL_2, MERL_1, and IOA3D outperform the average HRTF in all cases, followed by UDiPD and SupDEq, which are better than the average HRTF in at least half the cases. These are also the only five algorithms that are superior to the average HRTF with respect to all error measures at least for difficulty 100. While the remaining upsampling approaches all outperform the baseline methods Barycentric and SH, they are inferior to the average HRTF in more than half the cases. SYT_FSP_AE and Kalimotxo are the only two submissions that perform worse than the average HRTF for all auditory metrics, which might indicate an over-fitting of these algorithms to the LAP metrics. Overall, the best results are clearly obtained for LSD, which again might indicate an over-fitting to this specific metric.

B. EVALUATION FRAMEWORK

While the objective evaluation framework used in this challenge provides a solid foundation, additional refinements could enhance its effectiveness. One area of interest is the potential for generative models to introduce artefacts or 'hallucinations', particularly when upsampling from extremely sparse input data (e.g. using only 5 or 3 HRTFs). auralizations of moving sources can help to identify abrupt colouration artefacts caused by these models, and to this end, we have provided such auralizations as supplemental material.⁶

Another potential improvement to the evaluation framework is the development of a differentiable metric for assessing spatial smoothness in generated HRTFs. This would help ensure that reconstructed HRTFs exhibit natural continuity across spatial directions, avoiding perceptually implausible discontinuities. Such discontinuities can occur, for example, when sparse input data is directly used without proper spatial interpolation or smoothing (as illustrated by the white dots in Fig. 4). A spatial smoothness metric could serve as an additional validity check during model development. Furthermore, perceptual testing could help determine whether these discontinuities introduce audible artefacts. It is possible that

⁶[Online]. Available: <https://github.com/Audio-Experience-Design/LAPChallenge>

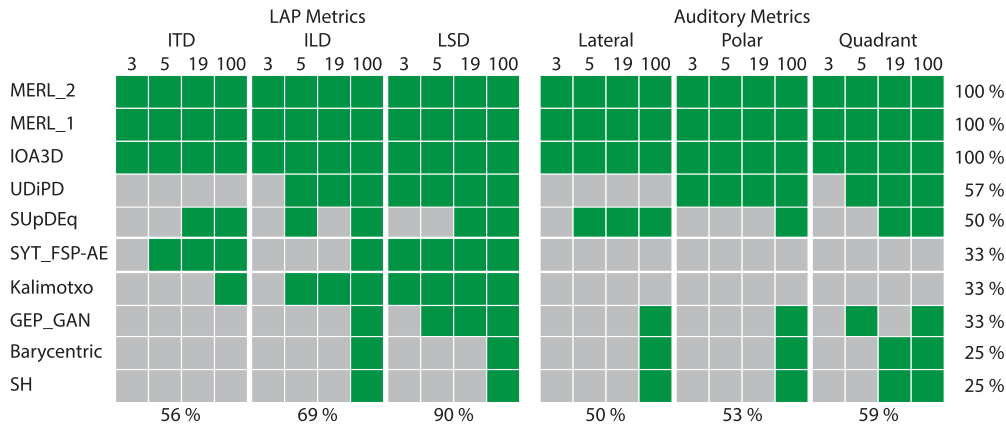


FIGURE 9. Comparison of upsampling algorithms. A green field denotes that an algorithm outperforms the SONICOM average HRTF. LAP metrics refer to Fig. 2, Auditory metrics to Fig. 8. The percentages denote in how many cases the average HRTF is outperformed. The baseline methods, Barycentric and SH, were excluded from the percentages per error measure given at the bottom of the figure.

such artefacts are perceived as more disruptive than larger, but spatially smoother, errors in magnitude. Future work could also explore extending such metrics to incorporate frequency and temporal continuity.

A final consideration for future evaluations is to avoid relying on a single metric in isolation, such as LSD, given their limitations. Instead, a more holistic approach, integrating multiple perceptually relevant measures, will be better suited to ensure a well-rounded assessment of HRTF upsampling quality.

By addressing these challenges and refining the evaluation framework, future iterations of the LAP challenge could provide even more robust insights into the effectiveness of different HRTF upsampling approaches.

VIII. CONCLUSION

In this paper, the findings from Task 2 of the LAP Challenge have been presented. The results demonstrate key differences between algorithmic and machine learning-based approaches, demonstrating that learning-based methods—particularly those using deep neural networks—generally outperformed traditional upsampling techniques on objective metrics at lower sparsity levels. However, the results also highlight that at these lower sparsity levels, learning-based approaches seem to tend towards learning the average HRTF rather than the ground truth reference.

The paper also shows that while objective metrics offer valuable insights into upsampling performance, they also underscore the need for perceptual validation through both auditory models (and subjective listening tests). Notably, discrepancies were presented between the objective evaluation metrics and the results of an auditory model, which is designed to capture perception through localisation accuracy. Additionally, the auralizations provided alongside this paper indicate the need for a smoothness metric, which is currently not captured in existing assessments of HRTF upsampling.

Ultimately, this paper demonstrates the significant impact of metric choice on the evaluation of HRTF upsampling methods, showing that different assessment criteria can lead to varying conclusions about performance. The paper highlights the need for improvements in the current evaluation practices, as the limitations of existing objective metrics become evident when compared with perceptually motivated assessments. A more comprehensive evaluation framework—incorporating perceptual models, smoothness metrics and psychoacoustically weighted measures is needed to ensure a robust evaluation that aligns with human perception.

ACKNOWLEDGMENT

The authors would like to extend our gratitude to all the teams who participated in the challenge and contributed their innovative solutions, as well as to the other LAP Challenge organisers for making this event possible. The authors are also grateful to the IEEE Signal Processing Society (SPS) for their generous support as our main sponsors. Finally, the authors owe a special thank you to U-Sound and Dreamwaves, our industrial partners, for sponsoring the award ceremony at the 32nd European Signal Processing Conference (EUSIPCO 2024).

REFERENCES

- [1] T. Lokki et al., “Application scenarios of wearable and mobile augmented reality audio,” in *Proc. Audio Eng. Soc. Conv.*, May 2004, pp. 618–639.
- [2] D. Vickers et al., “Involving children and teenagers with bilateral cochlear implants in the design of the BEARS (both EARS) virtual reality training suite improves personalization,” *Front. Digit. Health*, vol. 3, Nov. 2021, Art. no. 759723.
- [3] M. Geronazzo and S. Serafin, Eds., *Sonic Interactions in Virtual Environments* (Human-Computer Interaction Series). Berlin, Germany: Springer, 2023.
- [4] D. N. Zotkin, R. Duraiswami, and L. S. Davis, “Rendering localized spatial audio in a virtual auditory space,” *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, Aug. 2004.
- [5] M. Geronazzo et al., “Technical Report: SONICOM/IEEE listener acoustic personalisation (LAP) challenge - 2024,” *IEEE Signal Process. Society, Tech. Rep.*, 2024, doi: [10.36227/techrxiv.173153187.72930965v1](https://doi.org/10.36227/techrxiv.173153187.72930965v1).

- [6] R. Daugintis et al., "Listener acoustic personalisation challenge - LAP24: Head-related transfer function dataset harmonization," *Open J. Signal Process.*, pp. 1–16, 2025.
- [7] D. Poirier-Quinot et al., "HRTF performance evaluation: Methodology and metrics for localisation accuracy and learning assessment," in *Adv. in Fundam. and Appl. Res. on Spatial Audio*. London, U.K.: IntechOpen, May 2022, pp. 63–95.
- [8] V. Best et al., "Sound Externalization: A Review of recent research," *Trends Hear.*, vol. 24, Jan. 2020, Art. no. 2331216520948390.
- [9] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acust. United Acust.*, vol. 91, no. 3, pp. 409–420, May 2005.
- [10] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the authenticity of individual dynamic binaural synthesis," *J. Acoust. Soc. Amer.*, vol. 142, no. 4, pp. 1784–1795, Oct. 2017.
- [11] L. Picinali and B. F. G. Katz, "System-to-user and user-to-system adaptations in binaural audio," in *Sonic Interactions in Virtual Environments*, M. Geronazzo and S. Serafin, Eds. Berlin, Germany: Springer, Oct. 2022, pp. 121–144.
- [12] R. Barumerli et al., "A Bayesian model for human directional localization of broadband static sound sources," *Acta Acustica*, vol. 7, May 2023, Art. no. 12.
- [13] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. I: Stimulus synthesis," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 858–867, Feb. 1989.
- [14] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1983.
- [15] P. Stütt, L. Picinali, and B. F. G. Katz, "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning," *Sci. Rep.*, vol. 9, no. 1, Jan. 2019, Art. no. 1063.
- [16] E. M. Wenzel et al., "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 94, no. 1, pp. 111–123, Jul. 1993.
- [17] H. Møller et al., "Binaural technique: Do we need individual recordings?," *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, Jun. 1996.
- [18] Y. Kahana and P. A. Nelson, "Numerical modelling of the spatial acoustic response of the human pinna," *J. Sound Vib.*, vol. 292, no. 1, pp. 148–178, Apr. 2006.
- [19] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, "Perceptual attributes for the comparison of head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 140, no. 5, pp. 3623–3632, Nov. 2016.
- [20] S. Werner et al., "A summary on acoustic room divergence and its effect on externalization of auditory events," in *Proc. Int. Conf. Qual. Multimedia Experience*, Jun. 2016, pp. 1–6.
- [21] I. Engel et al., "The effect of generic headphone compensation on binaural renderings," in *Proc. Audio Eng. Soc. Int. Conf. Immersive Interactive Audio*, Mar. 2019, pp. 1–10.
- [22] M. Cuevas-Rodríguez et al., "Impact of non-individualised head related transfer functions on speech-in-noise performances within a synthesised virtual environment," *J. Acoust. Soc. Amer.*, vol. 149, no. 4, pp. 2573–2586, Apr. 2021.
- [23] I. Engel et al., "The SONICOM HRTF dataset," *J. Audio Eng. Soc.*, vol. 71, pp. 241–253, Jun. 2023.
- [24] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "Mesh2HRTF: An open-source software package for the numerical calculation of head-related transfer functions," in *Proc. Int. Cong. Sound Vib.*, Jul. 2015, pp. 1–8.
- [25] B. F. G. Katz, "Boundary element method calculation of individual head-related transfer function. I. rigid model calculation," *J. Acoust. Soc. Amer.*, vol. 110, no. 5, pp. 2440–2448, Nov. 2001.
- [26] P. Stütt and B. F. G. Katz, "Sensitivity analysis of pinna morphology on head-related transfer functions simulated via a parametric pinna model," *J. Acoust. Soc. Amer.*, vol. 149, no. 4, pp. 2559–2572, Apr. 2021.
- [27] M. Geronazzo et al., "Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization," *J. Audio Eng. Soc.*, vol. 67, no. 6, pp. 414–428, Jun. 2019.
- [28] D. Y. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2003, pp. 157–160.
- [29] B. F. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *J. Acoust. Soc. Amer.*, vol. 131, no. 2, pp. EL99–EL105, Feb. 2012.
- [30] C. Kim, V. Lim, and L. Picinali, "Investigation into consistency of subjective and objective perceptual selection of non-individual head-related transfer functions," *J. Audio Eng. Soc.*, vol. 68, no. 11, pp. 819–831, Dec. 2020.
- [31] F. Zagala, M. Noisternig, and B. F. Katz, "Comparison of direct and indirect perceptual head-related transfer function selection methods," *J. Acoust. Soc. Amer.*, vol. 147, no. 5, pp. 3376–3389, May 2020.
- [32] T. Carpentier et al., "Measurement of a head-related transfer function database with high spatial resolution," in *Proc. EAA Forum Acusticum, Eur. Congr. Acoust.*, Krakow, Poland, Sep. 2014, pp. 1–6.
- [33] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. Audio Eng. Soc. Conv.*, Feb. 2000, pp. 1–23.
- [34] D. N. Zotkin et al., "Fast head-related transfer function measurement via reciprocity," *J. Acoust. Soc. Amer.*, vol. 120, no. 4, pp. 2202–2215, Oct. 2006.
- [35] J.-G. Richter, G. Behler, and J. Fels, "Evaluation of a fast HRTF measurement system," in *Proc. Audio Eng. Soc. Conv.*, May 2016, vol. 140, Art. no. 9498.
- [36] X.-L. Zhong and B.-S. Xie, *Head-Related Transfer Functions and Virtual Auditory Display*. Plantation, FL, USA: InTech, Mar. 2014.
- [37] J. M. Arend, C. Pörschmann, S. Weinzierl, and F. Brinkmann, "Magnitude-corrected and time-aligned interpolation of head-related transfer functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3783–3799, 2023.
- [38] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *Proc. Audio Eng. Soc. Conf. Spatial Sound Reproduction*, Mar. 1999, pp. 319–329.
- [39] D. Poirier-Quinot and B. F. G. Katz, "The Anaglyph binaural audio engine," in *Proc. Audio Eng. Soc. Conv.*, May 2018, pp. 1–4.
- [40] M. Cuevas-Rodríguez et al., "3D tune-in toolkit: An open-source library for real-time binaural spatialisation," *PLoS One*, vol. 14, no. 3, Mar. 2019, Art. no. e0211899.
- [41] H. Gamper, "Head-related transfer function interpolation in azimuth, elevation, and distance," *J. Acoust. Soc. Amer.*, vol. 134, no. 6, pp. EL547–EL553, Dec. 2013.
- [42] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *J. Acoust. Soc. Amer.*, vol. 104, no. 4, pp. 2400–2411, Jun. 1998.
- [43] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional equalization of sparse head-related transfer function sets for spatial upsampling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1060–1071, Jun. 2019.
- [44] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *J. Audio Eng. Soc.*, vol. 69, no. 1/2, pp. 104–117, Feb. 2021.
- [45] I. Engel, D. F. M. Goodman, and L. Picinali, "Assessing HRTF preprocessing methods for Ambisonics rendering through perceptual models," *Acta Acustica*, vol. 6, Jan. 2022, Art. no. 4.
- [46] D. Fantini, M. Geronazzo, F. Avanzini, and S. Ntalampiras, "A survey on machine learning techniques for head-related transfer function individualization," *IEEE Open J. Signal Process.*, vol. 6, pp. 30–56, 2025.
- [47] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, "Autoencoding HRTFs for DNN based HRTF personalization using anthropometric features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019, pp. 271–275.
- [48] D. Yao et al., "An individualization approach for head-related transfer function in arbitrary directions based on deep learning," *JASA Exp. Lett.*, vol. 2, no. 6, Jun. 2022, Art. no. 064401.
- [49] Y. Ito et al., "Head-related transfer function interpolation from spatially sparse measurements using autoencoder with source position conditioning," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2022, pp. 1–5.
- [50] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs [head related transfer functions]," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2004, vol. 4, pp. iv–iv.
- [51] G. Kestler, S. Yadegari, and D. Nahamoo, "Head related impulse response interpolation and extrapolation using deep belief networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019, pp. 266–270.

- [52] Z. Jiang et al., “Modeling individual head-related transfer functions from sparse measurements using a convolutional neural network,” *J. Acoust. Soc. Amer.*, vol. 153, no. 1, pp. 248–259, Jan. 2023.
- [53] B. Tsui, W. A. P. Smith, and G. Kearney, “Low-order spherical harmonic HRTF restoration using a neural network approach,” *Appl. Sci.*, vol. 10, no. 17, Jan. 2020, Art. no. 5764.
- [54] H. Xuyi et al., “HRTF spatial upsampling in the spherical harmonics domain employing a generative adversarial network,” in *Proc. Conf. Digit. Audio Effects*, Sep. 2024, pp. 396–403.
- [55] H. Xuyi et al., “Head-related transfer function upsampling using an autoencoder-based generative adversarial network with evaluation framework,” *J. Audio Eng. Soc.*, pp. 1–13, 2025.
- [56] H. Xuyi et al., “A machine learning approach for denoising and upsampling HRTFs,” in *Proc. Eur. Signal Process. Conf.*, Sep. 2025, pp. 1–5.
- [57] J. C. A. Sánchez, L. Comanducci, M. Pezzoli, and F. Antonacci, “Towards HRTF personalization using denoising diffusion models,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2025, pp. 1–5.
- [58] A. O. T. Hogg, M. Jenkins, H. Liu, I. Squires, S. J. Cooper, and L. Picinali, “HRTF upsampling with a generative adversarial network using a gnomonic equiangular projection,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2085–2099, 2024.
- [59] P. Gutierrez-Parera et al., “Interaural time difference individualization in HRTF by scaling through anthropometric parameters,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2022, no. 1, pp. 1–19, May 2022.
- [60] T. McKenzie, D. T. Murphy, and G. Kearney, “Interaural level difference optimization of binaural ambisonic rendering,” *Appl. Sci.*, vol. 9, no. 6, Jan. 2019, Art. no. 1226.
- [61] P. Majdak, R. Baumgartner, and B. Laback, “Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization,” *Front. Psychol.*, vol. 5, Apr. 2014, Art. no. 319.
- [62] P. Majdak, “ARI HRTF database,” Jun. 2022. [Online]. Available: <http://www.kfs.oeaw.ac.at/hrtf>
- [63] R. Daugintis, R. Barumerli, L. Picinali, and M. Geronazzo, “Classifying non-individual head-related transfer functions with a computational auditory model: Calibration and metrics,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.
- [64] J. C. Middlebrooks, “Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency,” *J. Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1493–1510, Aug. 1999.
- [65] J. C. Middlebrooks, “Sound localization,” in *Handbook of Clinical Neurology*, vol. 129. Amsterdam, The Netherlands: Elsevier, 2015, pp. 99–116.
- [66] P. Lladó et al., “Spectral weighting of monaural cues for auditory localization in sagittal planes,” *Trends Hear.*, vol. 29, pp. 1–14, 2025.



ROBERTO BARUMERLI received the M.Sc. degree in 2016, and the Ph.D. degree in information and communication technology from the University of Padova, Padua, Italy, in 2021, focusing on computational auditory models for sound localisation. He was a Postdoctoral Researcher with the Acoustics Research Institute, Vienna, Austria. He is currently a Postdoctoral Researcher at the University of Verona, Italy. He specialises in experimental design, computational statistics, and auditory model development, applying these techniques to neuroscience, acoustics, and human-computer interaction. His research interests include the dynamics of auditory spatial perception through Bayesian inference and probabilistic modelling, integrating signal processing with behavioural and physiological data analysis. Up-to-date information on his research is available here: www.robertobarumerli.com.



RAPOLAS DAUGINTIS received the M.Sc. degree in acoustics and audio technology from Aalto University, Espoo, Finland, in 2021. He is currently working toward the Ph.D. degree with Audio Experience Design team, Dyson School of Design Engineering, Imperial College London, London, U.K. He is also investigating perceptual differences of head-related transfer functions and the relevance of binaural spatial audio personalisation in various extended reality scenarios. His work includes collaborations with members of the EU-funded SONICOM project, which aims to improve immersive audio experiences for extended reality applications. He also had some industry experience working as an Acoustics Consultant. His research interests include psychoacoustics, spatial audio, auditory & acoustic modeling, and extended reality.



KATARINA C. POOLE received the M.Sc. degree in 2017, and the Ph.D. degree in auditory neuroscience—investigating how the brain extracts auditory patterns through methodologies such as psychophysics, electrophysiology, and optogenetics from University College London, London, U.K. She is currently a Research Associate in immersive audio with Audio Experience Design Team, Dyson School of Design Engineering, Imperial College London, U.K. Her research interests include assessing numerical and perceptual differences between measured and synthetic head-related transfer functions. Her work includes collaboration with members of the EU-funded SONICOM project to improve immersive audio experiences as well as collaboration with UCL and the Eriksholm Research Centre to understand non-target spatial hearing in normal and hearing-impaired listeners.



FABIAN BRINKMANN received the M.A. degree in 2011 and the Ph.D. degree in 2019. He is currently a Postdoctoral Researcher with Audio Communication Group, Technische Universität Berlin, Berlin, Germany, where he also Leads the Binaural Processing and Perception Group. His research interests include the fields of signal processing and evaluation approaches for spatial audio. He is also a member of the AES, German Acoustical Society (DEGA), and the European Acoustics Association (EAA) Technical Committee for Psychological and

Physiological Acoustics.



AIDAN O. T. HOGG received the M.Eng. degree in electronic and information engineering and the Ph.D. degree from Imperial College London, London, U.K., in 2017 and 2022, respectively. He has worked in various engineering roles with Broadcom, Dialog Semiconductor, and Nuance Communications. He was a Research Associate in spatial audio and virtual reality with Imperial College London, Audio Experience Design Group, where he is currently an Honorary Research Associate. He is also a Lecturer of computer science and Co-Lead with Virtual, Immersive, Augmented and Binaural Audio Laboratory (VIABAL), Centre for Digital Music, Queen Mary University of London, London, U.K. His research interests include using deep learning to capture head-related transfer functions and, more generally, spatial acoustics, immersive audio, speaker diarisation, and statistical signal processing for audio applications. More information about current research projects can be found here: www.aidanhogg.uk.



LORENZO PICINALI received the M.Sc. degree in 2013 and the Ph.D. degree in 2010. He is currently a Professor of spatial acoustics and immersive audio with Imperial College London, London, U.K., where he also leads the Audio Experience Design Team (www.axdesign.co.uk). His research explores both the perceptual and computational aspects of acoustics alongside their practical applications. The aim is to understand how humans perceive spatial sound features, such as source locations and reverberation, and use this knowledge

to design algorithms and tools for creating virtual acoustic simulations. His research interests mainly include is to seamlessly blend real and virtual acoustic environments and ultimately enhancing hearing. He also works on eco-acoustic monitoring, designing autonomous recorders, and using audio to study the impact of human activity on remote ecosystems.



MICHELE GERONAZZO (Senior Member, IEEE) received the M.Sc. degree in 2009 and the Ph.D. degree in 2014. Between 2014 and 2021, he was an Assistant Professor with the University of Udine. He was also a Postdoctoral Researcher with Aalborg University, Copenhagen, Denmark. Since 2015, he has been a part of the organising committee of the IEEE VR Workshop on Sonic Interactions for Virtual Environments (chair of the 2018, 2020, and 2022 editions). He is currently an Associate Professor with the

University of Padova, Padua, Italy, and is also a part of the coordination unit of the EU-H2020 project SONICOM at Imperial College London. His research interests mainly include modeling and simulation of complex human-machine sonic interactions. He is also an Associate Editor for IEEE OPEN JOURNAL OF SIGNAL PROCESSING, *ACM Transactions on Applied Perception*, and the Editor of the Springer-Nature's Book *Sonic Interactions in Virtual Environments*. He was the co-recipient of six best paper/poster awards and co-author of more than ninety scientific publications.